ESCUELA SUPERIOR POLITÉCNICA DEL LITORAL

Facultad de Ingeniería en Electricidad y Computación

SISTEMA PARA EL ANÁLISIS Y PROCESAMIENTO DE LOS LOGS DE LOS SERVIDORES DE RED DE LA FACULTAD DE INGENIERÍA EN ELECTRICIDAD Y COMPUTACIÓN (FIEC) DE LA ESPOL USANDO HADOOP

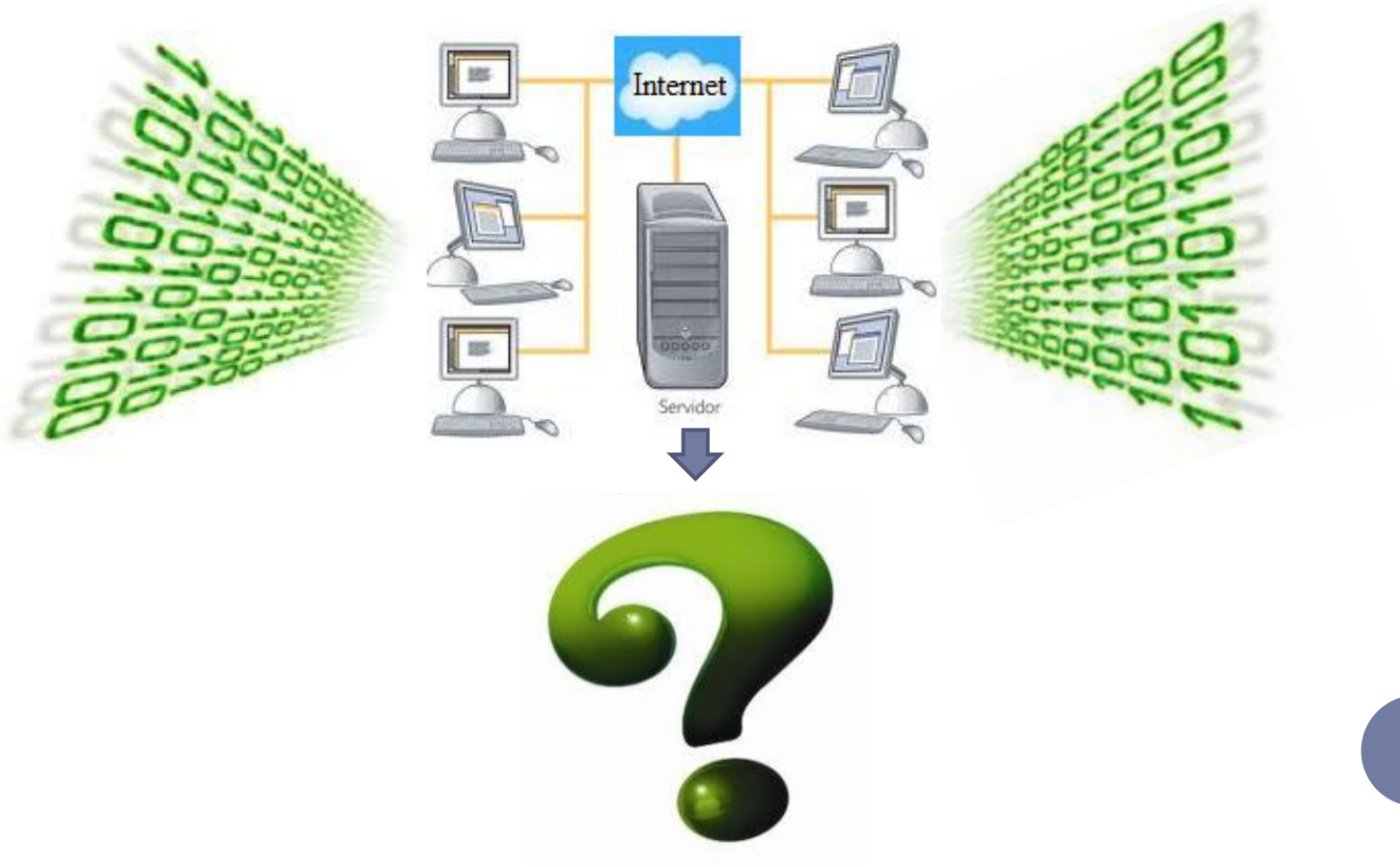
INTEGRANTES:

Eddy Roberto Espinosa Daquilema

Josué Jefferson Guartatanga Robayo

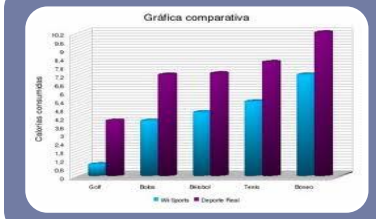
Problema a resolver

Datos vs. Información

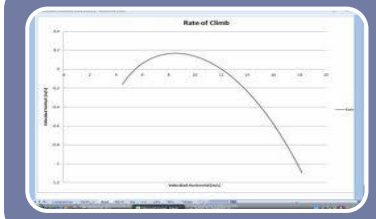


JUSTIFICACIÓN

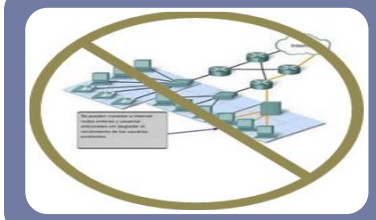
Existen herramientas con precios altos



No muestran resultados deseados



Rendimiento deficiente / tiempo extenso



No son distribuidas ni escalables

Alcance

Información de los servidores: Cedro(HTTP), Ceibo(Maillog) y Palma(Samba).

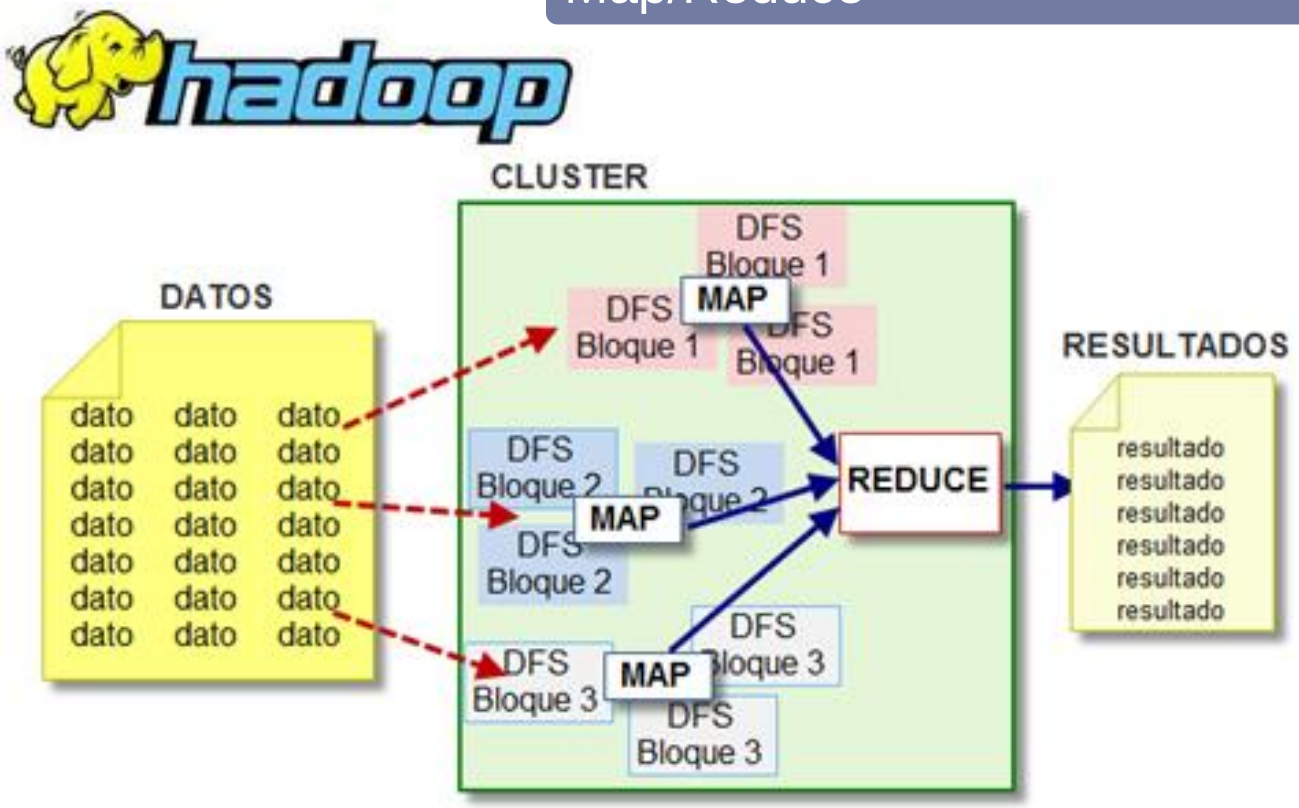
Visitas/Recursos/Navegadores

Correos

Acceso de usuarios/Recursos

Hadoop

- Plataforma -> aplicaciones escalables
- Desarrollador enfoca -> lógica de negocio
- Corre en cluster / (HDFS)
- Map/Reduce

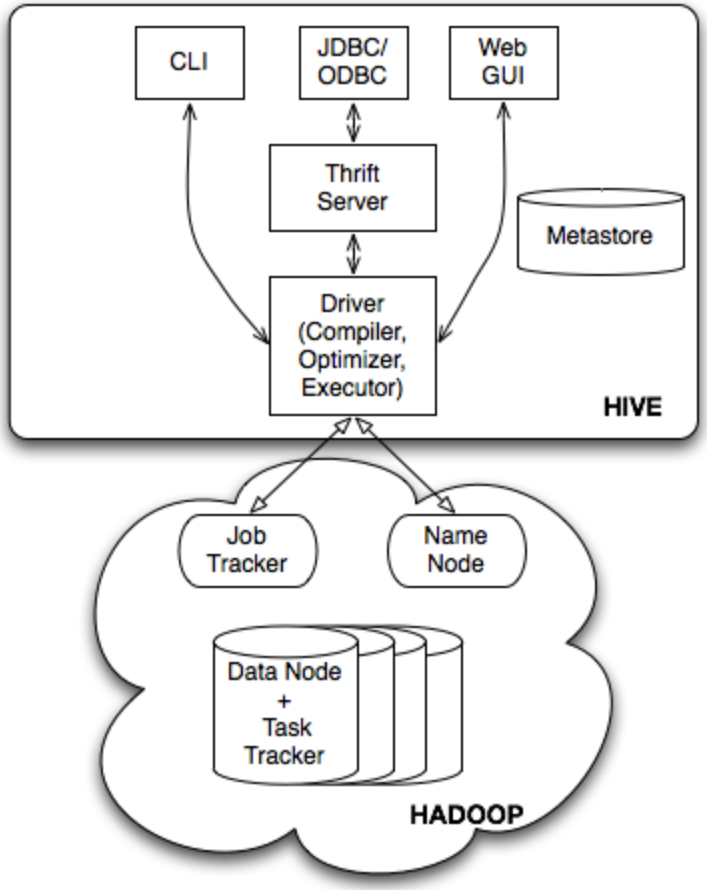


HIVE

Infraestructura para data warehouse que provee sumarización de datos y soporta Ad-hoc queries

Aplicaciones

- Logs
- Minería Datos
- Indexación documentos
- Inteligencia Negocios
- Modelamiento predictivo y Prueba de hipótesis



Formato de Logs

Cedro posee un Servidor Apache HTTP

```
67.195.112.238 - - [22/Nov/2009:04:13:29 -0500] "GET
/resources/materias/licred/FIEC06221_implementacion_soporte_xp_pr
of.pdf HTTP/1.0" 200 51590 "-" "Mozilla/5.0 (compatible; Yahoo!
Slurp/3.0; HTTP://help.yahoo.com/help/us/ysearch/slurp)"
```

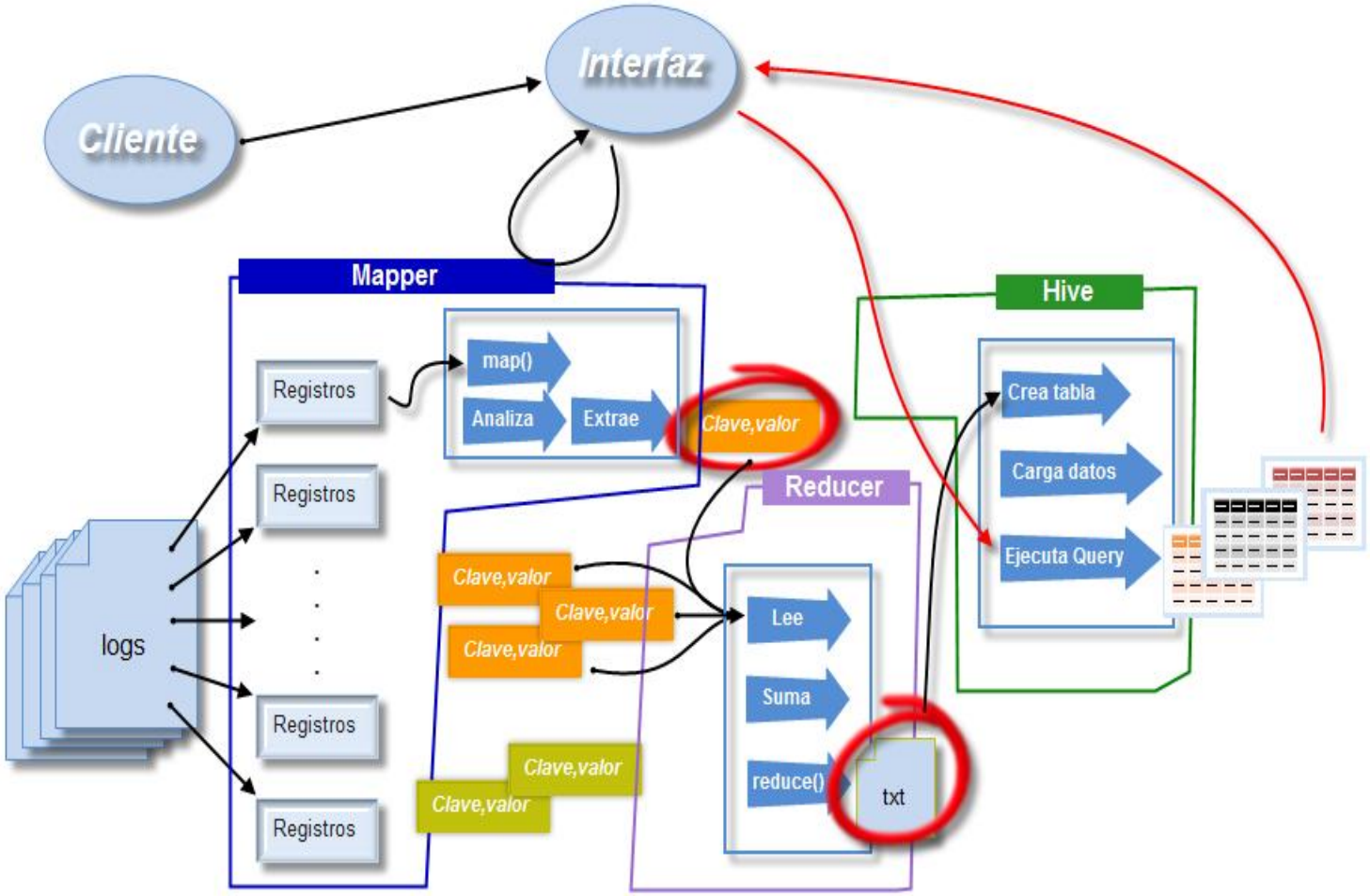
Ceibo genera archivos tipo Maillog

```
Nov 9 00:21:16 ceibo milter-greylist: nA95LGs1020166: addr
mail.periodicourbano.com[200.125.135.132] from
<nobody@mail.periodicourbano.com> to <amera@fiec.espol.edu.ec>
delayed for 00:05:00 (ACL 300)
```

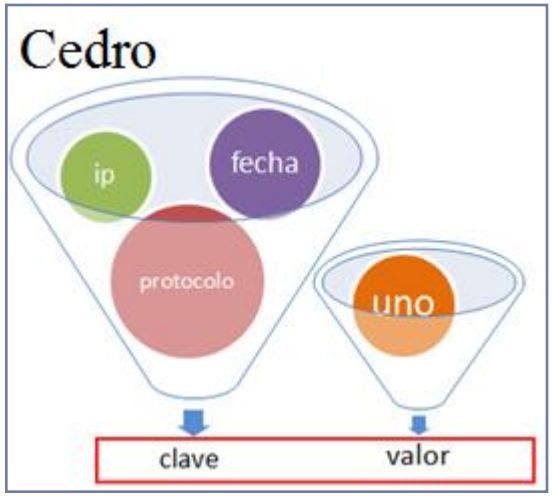
Palma posee un servidor Samba.

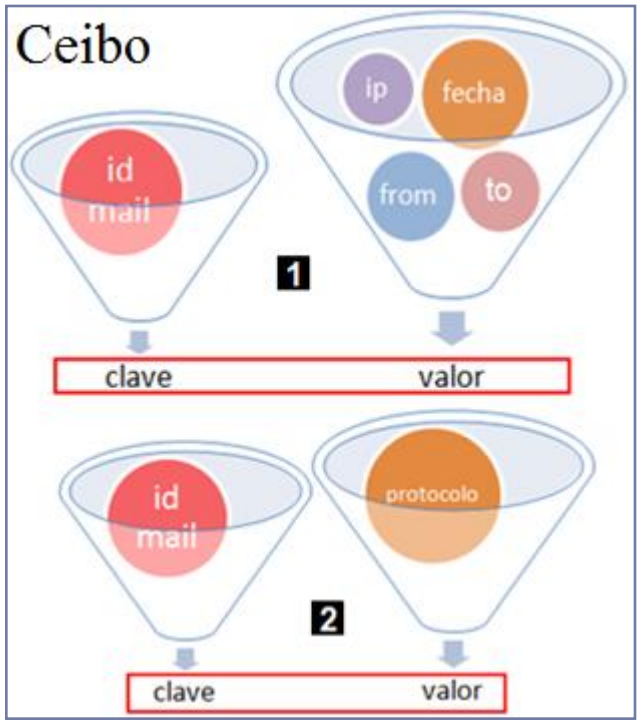
```
[2009/09/16 12:46:56, 1]
smbd/service.c:make_connection_snum(1033)
  wrks126-170fiec (200.9.176.170) connect to service apincay initially
as user apincay (uid=6828, gid=501) (pid 5491)
[2009/09/17 13:52:46, 2] smbd/open.c:open_file(391) opreciad opened
file Internet Download Manager/IDMan.exe read=Yes write=No
(numopen=11)
```

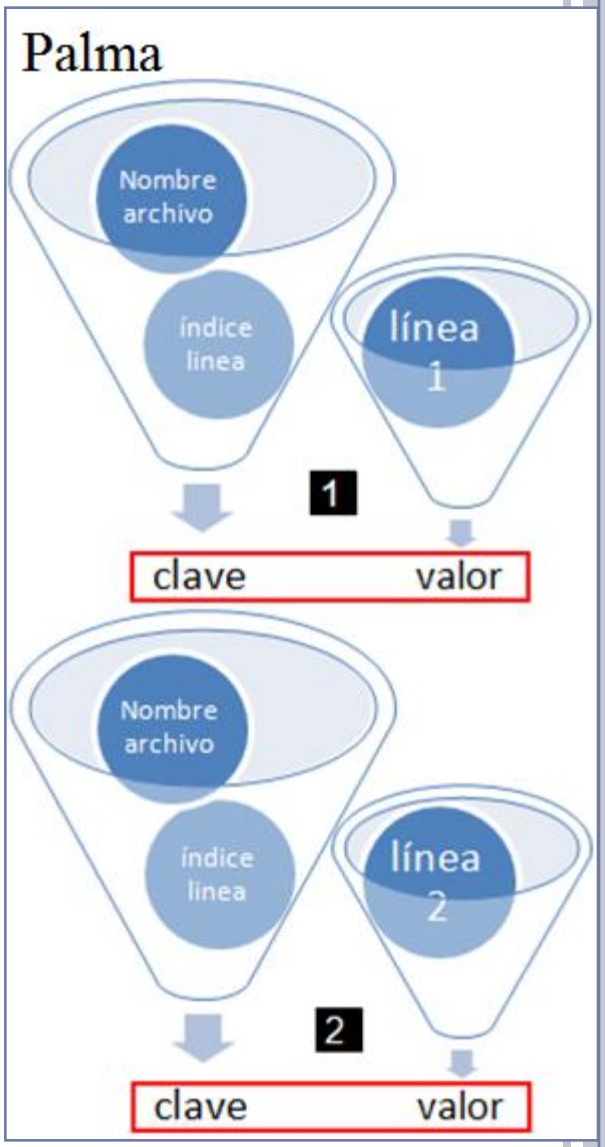
Diseño de la solución



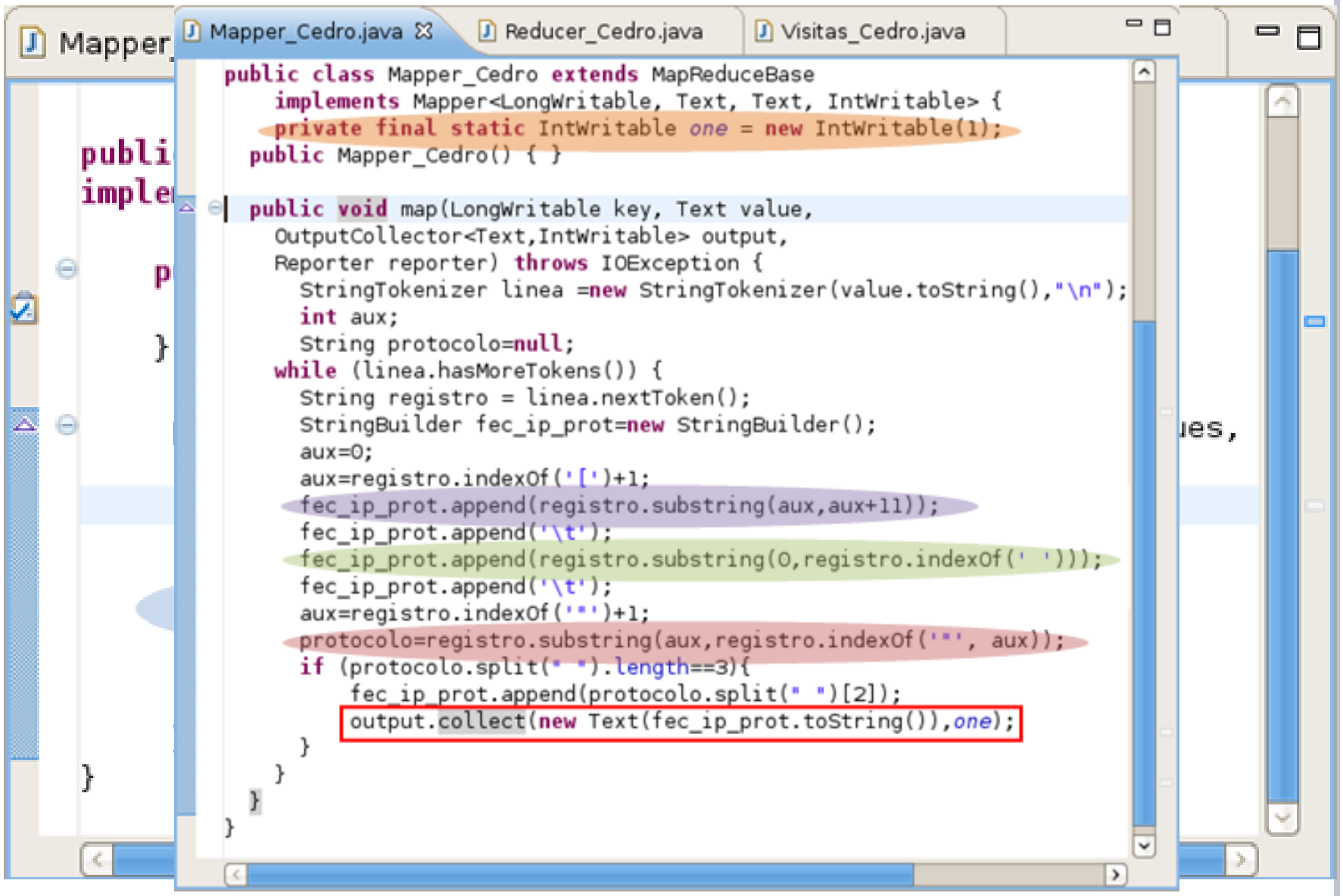
Claves y Valores







Código Base: Mapper/Reducer de Cedro

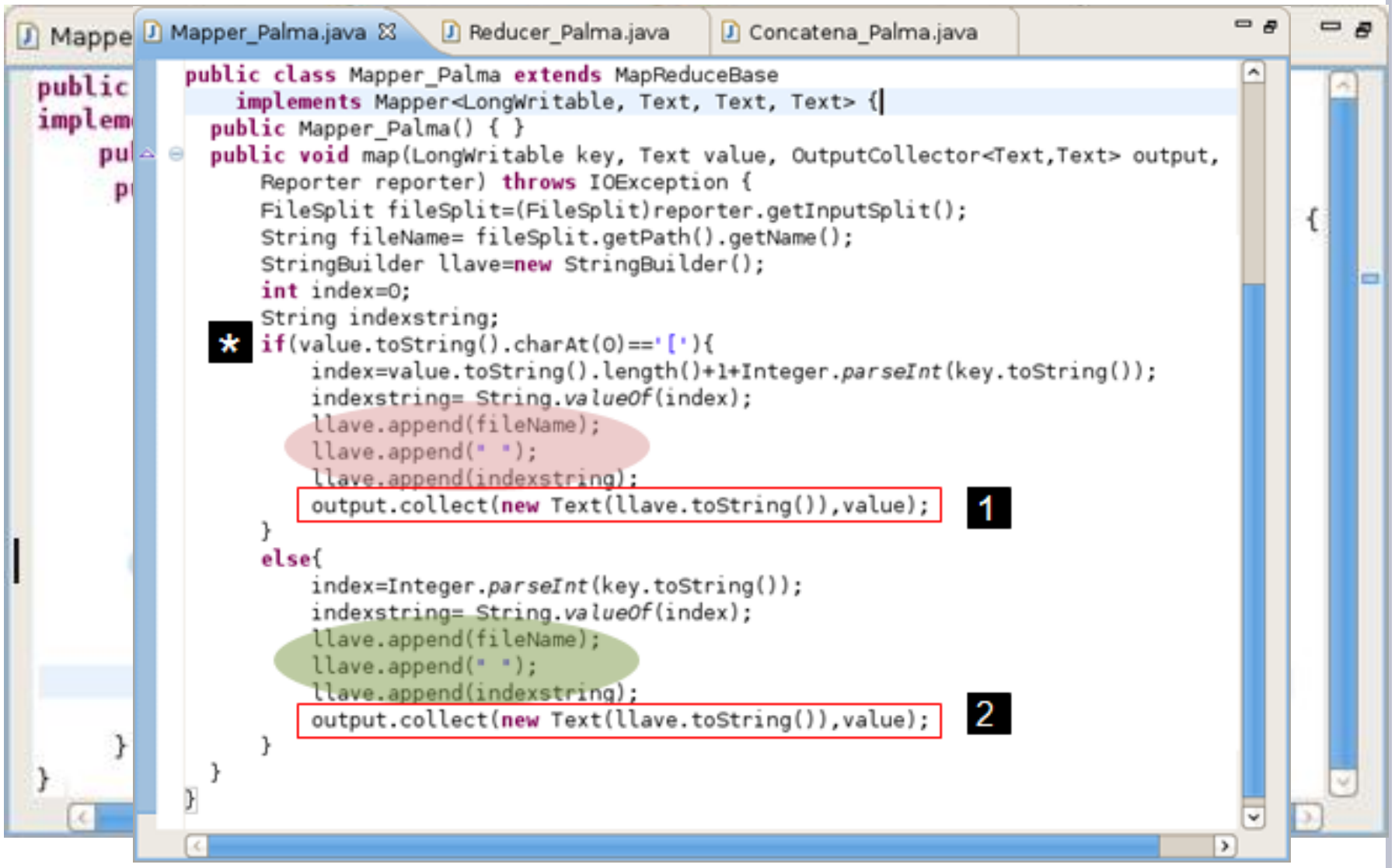


CÓDIGO BASE: MAPPER/REDUCER DE CEIBO

Mapper_Ceibo.java | Reducer_Ceibo.java | Correos_Ceibo.java

```
identificador=registro.indexOf("milter-greylist: ");
if(identificador!=-1){
    identificadormail=new Text(registro.substring(identificador+17,identificador+31));
    int indice=registro.indexOf(": addr ");
    if (indice!=-1){
        fecha_ip_from_to.append(registro.substring(0,6));
        fecha_ip_from_to.append('\t');
        indice=registro.indexOf('[',indice+7);
        if(indice!=-1){
        fecha_ip_from_to.append(registro.substring(indice+1, registro.indexOf(']',indice+1)));
        fecha_ip_from_to.append('\t');
        indice=registro.indexOf(" from <");
         if (indice!=-1){
            fecha_ip_from_to.append(registro.substring(indice+7,registro.indexOf('>',+indice)));
            fecha_ip_from_to.append('\t');
            indice=registro.indexOf(" to <");
            fecha_ip_from_to.append(registro.substring(indice+5,registro.indexOf('>',+indice)));
            output.collect(identificadormail,new Text(fecha_ip_from_to.toString())); 1
         }
        }
    }
}
else{
    identificador= registro.indexOf("sendmail[");
    if(identificador!=-1){
        int indice= registro.indexOf("]: ", identificador);
        if(indice!=-1){
        identificadormail=new Text(registro.substring(indice+3, indice+17));
        indiceprotocolo=registro.indexOf(", proto=", indice);
         if(indiceprotocolo!=-1){
        Protocolo=registro.substring(indiceprotocolo+8,registro.indexOf(", ",indiceprotocolo+9));
        output.collect(identificadormail,new Text(Protocolo)); 2
         }
```

Código Base: Mapper/Reducer de Palma



TIEMPOS

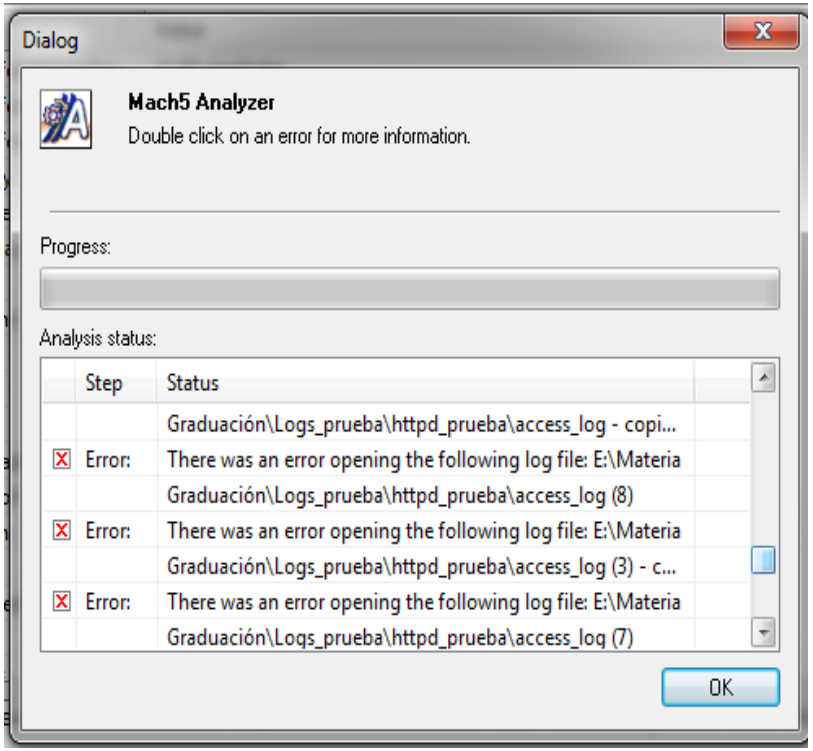
Tiempo (minutos)

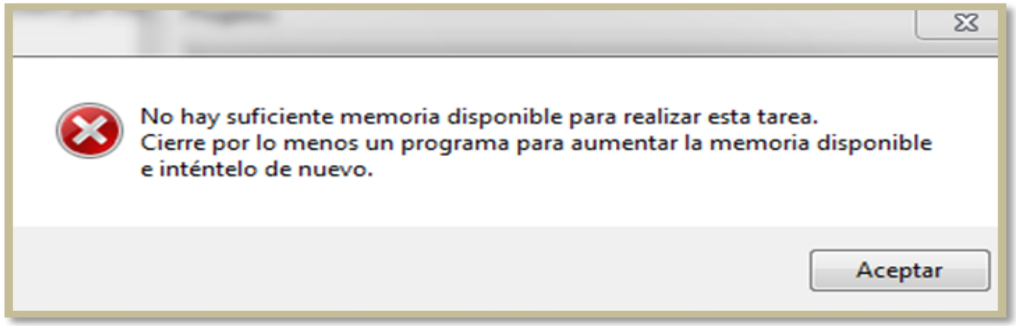
# prueba	tiempo (min)								
	carga 37.2 MB / 0.036 GB								
	Palma Concatena			Palma Accesos			Palma Recursos		
	Nodos								
	5	10	15	5	10	15	5	10	15
1	6,783	1,933	1,367	0,467	0,350	0,267	0,417	0,620	0,333
2	6,833	1,967	1,983	0,450	0,620	0,317	0,533	0,220	0,350
3	7,017	3,020	1,683	0,517	0,200	0,300	0,550	0,240	0,350
4	7,683	3,517	1,683	0,550	0,210	0,333	0,533	0,220	0,350
5	6,833	4,350	1,267	0,517	0,270	0,300	0,433	0,600	0,300
6	6,383	4,417	2,050	0,717	0,220	0,317	0,517	0,220	0,317
7	9,683	5,667	1,633	0,483	0,210	0,300	0,450	0,630	0,400
8	6,167	4,317	1,483	0,467	0,600	0,283	0,450	0,240	0,300
9	8,633	4,267	1,283	0,433	0,590	0,417	0,500	0,680	0,317
10	6,350	5,367	1,333	0,517	0,230	0,317	0,517	0,200	0,317
promedio	**7,237**	**3,882**	**1,577**	**0,512**	**0,350**	**0,315**	**0,490**	**0,387**	**0,333**
Var. Std.	**1,124**	**1,273**	**0,281**	**0,081**	**0,180**	**0,040**	**0,048**	**0,213**	**0,030**

Aplicación no Distribuida

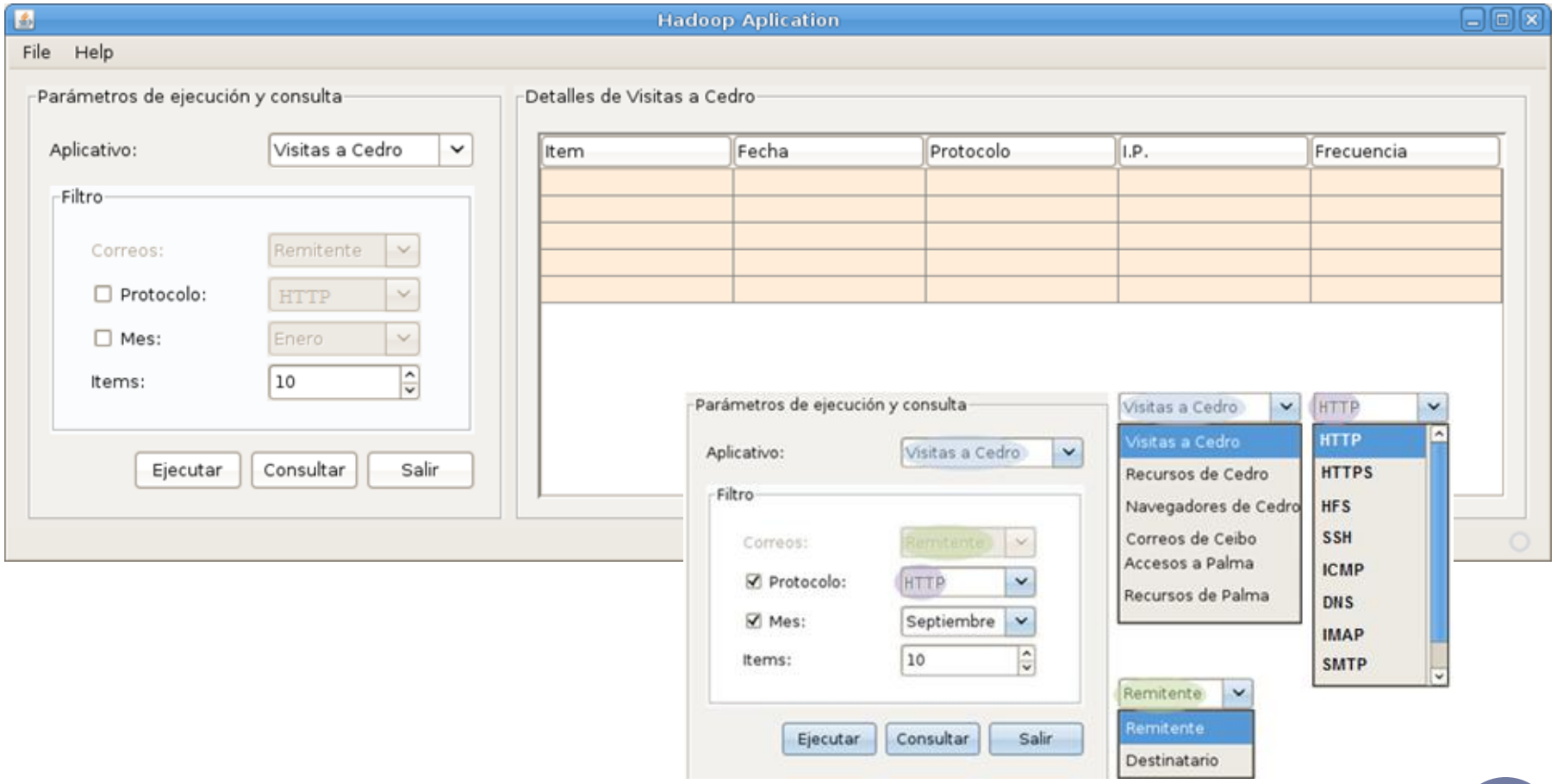
Carga

- 10 GB

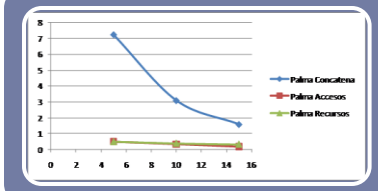




Aplicación Distribuida



Conclusiones



Estable entre 5 y 10 nodos con carga 1GB.

Procesamiento secuencial es extenso con archivos de gran tamaño.

El modelo distribuido es una solución escalable y la tolerancia a fallos se puede controlar.

Recomendaciones

Asignar suficiente espacio en disco para la instalación.

Familiarizarse con el entorno que ofrece Hadoop y su aplicación Hive.

Fin

MUCHAS GRACIAS !!!